Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations

Year: 2023 | Citations: 699 | Authors: Hakan Inan, K. Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer

---

Abstract

We introduce Llama Guard, an LLM-based input-output safeguard model geared towards Human-AI conversation use cases. Our model incorporates a safety risk taxonomy, a valuable tool for categorizing a specific set of safety risks found in LLM prompts (i.e., prompt classification). This taxonomy is also instrumental in classifying the responses generated by LLMs to these prompts, a process we refer to as response classification. For the purpose of both prompt and response classification, we have meticulously gathered a dataset of high quality. Llama Guard, a Llama2-7b model that is instruction-tuned on our collected dataset, albeit low in volume, demonstrates strong performance on existing benchmarks such as the OpenAI Moderation Evaluation dataset and ToxicChat, where its performance matches or exceeds that of currently available content moderation tools. Llama Guard functions as a language model, carrying out multi-class classification and generating binary decision scores. Furthermore, the instruction fine-tuning of Llama Guard allows for the customization of tasks and the adaptation of output formats. This feature enhances the model's capabilities, such as enabling the adjustment of taxonomy categories to align with specific use cases, and facilitating zero-shot or few-shot prompting with diverse taxonomies at the input. We are making Llama Guard model weights available and we encourage researchers to further develop and adapt them to meet the evolving needs of the community for AI safety.